INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

VOL 04. ISSUE 02 . (2026)

http://ijirjournal.com

Retrieval-Augmented Explainable Fake Review Detection with Evidence-Grounded Indicators

Riley Lin [1], Liam Cooper [1], Thomas Yoon [2*]

[1] School of Information Studies, Syracuse University, USA

[2] Adobe Research, Adobe Inc., USA

Corresponding author: *Thomas.Yoon19842@gmail.com*

Abstract.

Fake review detectors must be accurate and able to justify their decisions to moderators, sellers and regulators. Knowledge-enhanced detectors built on large language models (LLMs) have shown that curated fraud knowledge and step-by-step rationales improve both accuracy and interpretability, but they rely on GPU fine-tuning and generative inference. This paper presents GRADE (Grounded Retrieval-Augmented Detection with Explanations), a lightweight framework in which every decision is grounded in three verifiable sources of evidence: labelled precedent reviews retrieved by a hybrid sparse-dense retriever over pretrained language model embeddings, a knowledge base of 25 executable fraud indicators (linguistic, sentiment and behavioural) calibrated on the training data, and a sparse lexical model that exposes the n-grams driving the decision. An additive stacker combines the three channels, so the log-odds of each verdict decompose exactly into channel contributions and the explanation is faithful by construction. On two public benchmarks, human-written deceptive hotel reviews and machine-generated Amazon reviews, GRADE reaches 90.0% and 95.2% accuracy (AUC-ROC 96.7% and 99.0%), matching or exceeding the strongest non-retrieval baselines; it is up to 5.3 F1 points more accurate when only a few hundred labelled reviews are available, transfers across product categories with up to 8.7 F1 points more than text-only models, and runs at 54–150 ms per review on a single CPU core. Perturbation tests show that the evidence cited by GRADE, obtained at no extra cost, approaches the faithfulness of a LIME surrogate that needs 400 additional model evaluations per review and is three to nine times more faithful than random evidence at the same budget.

Keywords: *fake review detection; retrieval-augmented classification; explainable AI; evidence grounding; hybrid retrieval; faithfulness*

1. Introduction

Online reviews shape a large share of consumer spending, and the incentive to manipulate them is correspondingly strong. Field studies on Yelp found that a restaurant's propensity to post fraudulent reviews rises sharply when its reputation is weak or competition intensifies [1], and surveys document a steady escalation from crowd-sourced human deception to automatically generated text [2], [3]. Platforms therefore operate detection systems at scale, and their decisions affect sellers and users directly; a detector that cannot say *why* a review was flagged is difficult to

audit, to contest, and to maintain as fraud tactics drift. Detectors have moved through three generations. Early work combined hand-crafted linguistic and behavioural features with linear classifiers [4]–[6]; neural models then learned representations directly from text [7], [8]; and graph models exploited the relational structure between reviewers, products and reviews [9]. Most recently, large language models adapted with low-rank adapters [10] and prompted to reason step by step [11] have combined high accuracy with natural-language rationales, and knowledge-enhanced variants have shown that injecting curated fraud indicators improves both detection and interpretability (Section II-A). Operating such detectors at platform scale adds three requirements: inference should run on commodity hardware, newly labelled fraud should become usable without re-training a large model, and the rationale shown to a moderator should be verifiable against the computation that produced the verdict [12], [13]. Retrieval augmentation offers a different route to the same goals. Instead of storing task knowledge in model weights, retrieval-augmented systems, of which retrieval-augmented generation (RAG) is the best-known form, retrieve relevant evidence from an external index at inference time and condition the prediction on it [14]–[16]. For in-context learning, retrieving examples that are semantically similar to the query has been shown to be far more effective than random examples [17]. Two properties make this attractive for fraud detection: the evidence that supports a decision is explicit and can be displayed, and the evidence base can be updated by appending newly labelled reviews to the index. Knowledge-enhanced prompting applies the first property to curated fraud indicators; this paper asks whether a detector can be made explainable *by construction* if all of its evidence is retrieved, measured or weighted transparently. We present GRADE (Grounded Retrieval-Augmented Detection with Explanations), a retrieval-augmented classifier whose explanations are traceable to their evidence. GRADE grounds every decision in three channels: (i) an *exemplar channel* that retrieves the k most similar labelled reviews with a hybrid retriever combining BM25 and dense embeddings from a pretrained MiniLM sentence encoder, and converts their labels into a similarity-weighted vote; (ii) a *knowledge channel* that evaluates a knowledge base of 25 executable fraud indicators, organised into linguistic, sentiment and behavioural categories, whose direction and strength are calibrated on the training data and which are surfaced by measurement rather than by textual similarity; and (iii) a *lexical channel* that scores the review with a sparse n-gram model and exposes the phrases responsible. A logistic stacker combines the three channel scores, so the final log-odds decompose additively into channel contributions. Because the explanation lists the actual inputs of the decision function, removing the cited evidence and re-running the pipeline is a direct test of faithfulness, which we conduct against LIME and random-evidence controls.

The contributions of this work are as follows.

- A retrieval-augmented detection framework whose three evidence channels (retrieved labelled precedents, calibrated knowledge indicators and lexical cues) are combined additively, so that verdicts can be read off channel by channel.
- An executable fraud-indicator knowledge base that turns natural-language fraud knowledge into measurable functions, and a measurement-based selection rule that surfaces only the indicators a review actually exhibits, compared directly against similarity-based knowledge retrieval.
- A grounded explanation generator and an end-to-end perturbation-based faithfulness protocol: removing the cited evidence, which covers only about a tenth of the tokens of a review, lowers the model's confidence by 0.30 on average and flips 35% of the verdicts on OpSpam, matching a post hoc LIME surrogate on that corpus at zero overhead and exceeding a random-evidence control by an order of magnitude.
- A comprehensive evaluation on human-written deception (1,600 hotel reviews) and machine-generated reviews (40,432 Amazon reviews): ablations, retrieval-mode and k sensitivity, data

efficiency and index growth, transfer across polarity, category and dataset, and CPU-only efficiency.

II. Related Work

A. Fake Review Detection

Ott et al. introduced the first gold-standard corpus of deceptive hotel reviews and showed that n-gram features with a support vector machine detect crowd-sourced deception far better than human judges, who perform near chance [4]. Behavioural signals such as rating deviation, burstiness and reviewer activity were found to be highly indicative of spam on Yelp [5], and collective models that bridge review networks and metadata improved detection further [6]. A newer threat is machine-generated text: Salminen et al. showed that GPT-2-generated product reviews fool human readers but not classifiers [18]. Large language models have recently advanced the field considerably. Chen et al. [19] proposed KE-MLLM, which adapts LLaMA-3-8B with LoRA, retrieves curated fraud indicators into a knowledge-enhanced prompt, fuses temporal and user-level behavioural signals through cross-attention, distils knowledge from a 70B teacher and generates chain-of-thought rationales; it reaches an F1-score of 94.3% on YelpChi, and its explanations agree with expert annotations in 89.5% of cases. Two of its findings directly inspired the present work: fraud knowledge organised into linguistic, sentiment and behavioural categories improves detection substantially (removing its knowledge base costs 3.2 F1 points), and decisions become auditable when the rationale walks through these categories step by step. GRADE extends this knowledge-enhanced design to a retrieval-augmented setting in which the knowledge is executable and the decision is an additive function of verifiable evidence (Section III-C).

B. Retrieval-Augmented Methods and Faithful Explanations

Retrieval augmentation conditions a model on documents fetched from an external corpus: REALM pre-trains a language model jointly with a latent retriever [20], RAG combines a dense retriever with a sequence-to-sequence generator [14], and in-context RALM shows that simply prepending retrieved documents to a frozen language model yields large gains without any training [15]. GRADE's exemplar channel consumes the retrieved items through an interpretable weighted vote rather than an opaque reader, so that the precedents can be shown to the user as evidence. On the explanation side, the XAI literature distinguishes post hoc explanations of black-box models from models that are interpretable by design [21]. LIME fits a local surrogate around an instance [22]; attention weights were shown not to be reliable indicators of what drives a prediction [12]. Jacovi and Goldberg argue that explanations must be evaluated for *faithfulness*, i.e., whether they reflect the model's actual reasoning, rather than only for plausibility [13], and ERASER operationalises this with comprehensiveness and sufficiency metrics computed by erasing or retaining the cited rationale [23]. GRADE is designed so that such erasure tests are passed by construction, and Section VI applies them end-to-end. Neural-symbolic dual-indexing provides an instructive retrieval design that combines structured access paths with broad semantic coverage in retrieval-augmented systems [31].

III. The GRADE Framework

A. Problem Statement and Overview

Let $D = \{(x_i, y_i)\}$ be labelled reviews, $x_i = (t_i, m_i)$ the text with optional metadata (here the star rating) and $y_i \in \{0, 1\}$ the genuine/fake label. We seek the log-odds $f(x)$ of the fake class together with an explanation $e(x)$ that is *grounded* (every cited item is a span of the review, a measured property of it, or a retrieved labelled review) and *faithful* (removing cited items moves $f(x)$ away from the verdict). Fig. 1 summarises the framework. GRADE uses pretrained sentence embeddings for evidence retrieval and a discriminative evidence stacker for classification; it does not employ generative language-model inference. An incoming review is encoded once with the MiniLM encoder and then processed by three evidence channels. The exemplar channel retrieves the k most similar labelled reviews from the index and turns their labels into a similarity-weighted

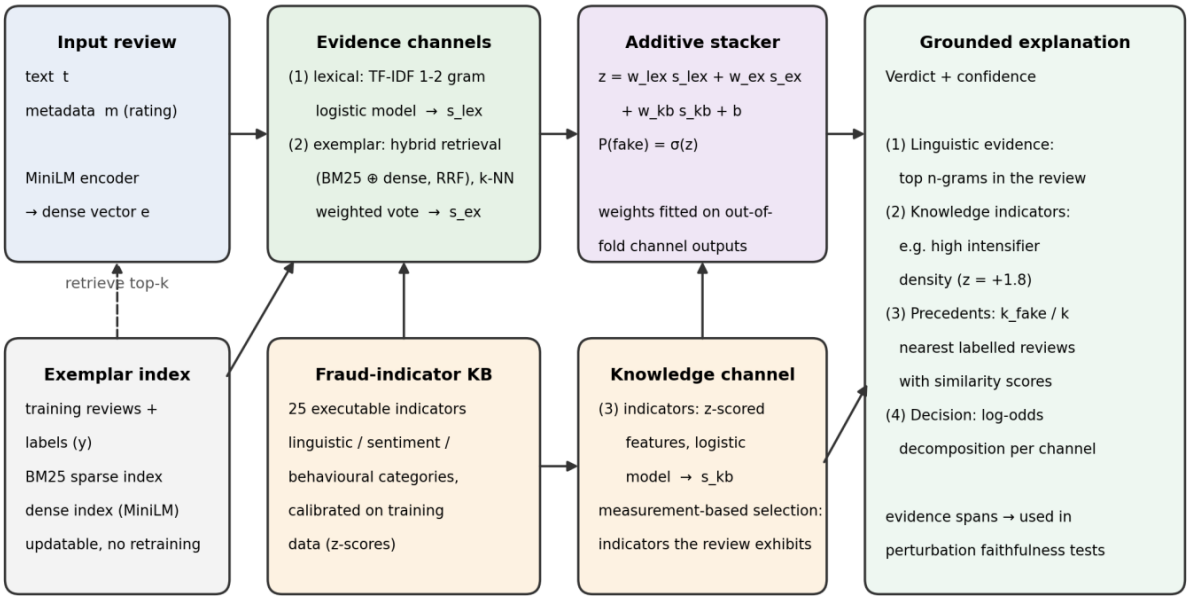


Figure 1Overview of GRADE. A review is encoded once; the exemplar channel retrieves labelled precedents with a hybrid BM25/dense retriever, the knowledge channel evaluates the calibrated fraud-indicator knowledge base, and the lexical channel scores n-grams. An

vote s_ex. The knowledge channel evaluates the fraud-indicator knowledge base on the review, standardises each indicator against the training distribution and scores the resulting vector with a logistic model, giving s_kb. The lexical channel scores the review with a sparse TF-IDF n-gram logistic model, giving s_lex. An additive stacker combines the three scores into the final log-odds z, and the explanation module assembles a grounded rationale from the components that produced z. Every component is deliberately linear in its evidence so that contributions can be attributed exactly; the non-linearity lies in retrieval and in the encoder, which expose their intermediate results (neighbours and similarities) in the explanation.

B. Hybrid Exemplar Retrieval

The index stores every training review together with its label, its BM25 term statistics and its dense embedding. Every review is embedded once with all-MiniLM-L6-v2, a six-layer transformer distilled from larger models [24] and trained as a sentence encoder with a contrastive objective [25]; mean pooling and $\ell 2$ normalisation give a 384-dimensional vector e(t) with cosine similarity $\cos(t, d) = e(t) \cdot e(d)$. For a query review t, the dense retriever ranks indexed reviews by cosine, the sparse retriever by BM25 [26] ($k_1 = 1.5$, $b = 0.75$, implemented as a sparse matrix product), and the two rankings are merged with reciprocal rank fusion [27], $RRF(d) = \Sigma_r 1/(\kappa + rank_r(d))$ with $\kappa = 60$, which is robust to the different scales of the two scores. The top-k fused results form the retrieved set R(t). BM25 surfaces reviews that reuse the same rare words and phrases, characteristic of templated spam, whereas dense similarity surfaces reviews with the same meaning expressed differently, characteristic of paraphrased or generated spam. We analyse the two retrievers and their fusion separately in Section V.

The retrieved labels are aggregated into a signed, similarity-weighted vote

$$s_{ex}(t) = \Sigma_{d \in R(t)} \omega_d (2y_d - 1) / \Sigma_{d \in R(t)} \omega_d, \quad \omega_d = \max(\cos(t, d), 0.05), \quad (1)$$

which lies in $[-1, 1]$ and is positive when the precedents are predominantly fake. Equation (1) is itself an explanation: it states how many of the k closest labelled reviews are fake and how similar they are, and the top precedents are displayed verbatim in the rationale.

C. Executable Fraud-Indicator Knowledge Base

The knowledge base is built upon the fraud-knowledge design of KE-MLLM [19]. We adopt its organisation of fraud knowledge into linguistic patterns, sentiment anomalies and behavioural red flags, which also structures its step-by-step rationales, and we implement its representative knowledge items as executable indicators, each consisting of a natural-language description and a function that measures the indicator on a review: excessive superlatives (L1), generic descriptions lacking product-specific details (L5–L7, L17) and rating–text sentiment mismatch (B2). Its rating-deviation signal is adapted as rating extremity (B1), measured against the scale midpoint because product-level average ratings are not available in the public corpora. These seven items are marked † in Table I. The remaining 18 indicators extend the same categories with further linguistic cues, such as punctuation and capitalisation, pronoun and adverb use, lexical diversity, word and review length, sentence-length variability, repetition, past-tense narration and function-word rate, and with sentiment cues computed by VADER [28]: sentiment extremity, the shares of positive and negative wording, and their minimum, which captures balanced evaluations. Burst and account-age red flags can be added as further items wherever a platform exposes timestamps and account histories.

Rather than fixing the polarity of each indicator by hand, GRADE calibrates the knowledge base on the training split. Every indicator value is standardised to a z-score using the training mean and standard deviation, and a regularised logistic model with coefficients β is fitted on the standardised vector, so that the knowledge score is

$$s_kb(t) = \beta_0 + \Sigma_j \beta_j z_j(t), \quad z_j(t) = (f_j(t) - \mu_j) / \sigma_j. \quad (2)$$

The sign of β_j encodes the learned direction of the indicator in the current domain (for example, first-person pronouns are a deception cue in crowd-sourced hotel reviews but not necessarily in generated product reviews), and the product $\beta_j z_j(t)$ is the contribution of indicator j to the decision for review t. An indicator is *triggered* for a review when its contribution points toward the predicted class and $|z_j(t)| \geq 0.5$; triggered indicators are reported in the explanation with their category, direction and magnitude, e.g., "high density of superlatives/intensifiers (z = +1.8)". This measurement-based selection adapts the knowledge retrieval of the adopted design, which ranks knowledge items by the cosine similarity between their descriptions and the review, an effective way to direct an LLM's attention through its prompt. With executable items, selection can follow what the review measurably exhibits rather than what it talks about, and every surfaced item carries a quantitative deviation score; Section V-B compares the two rules.

D. Lexical Channel

The lexical channel is a logistic regression on sub-linear TF-IDF features over word unigrams and bigrams (minimum document frequency 2). Its score $s_lex(t) = \theta_0 + \Sigma_g \theta_g x_g(t)$ is linear in the n-gram weights $x_g(t)$, and the explanation lists the n-grams present in the review with the largest contribution $\theta_g x_g(t)$ toward the predicted class. Sparse n-gram models remain among the strongest in-domain detectors of crowd-sourced deception [4].

E. Additive Stacking

The three channel scores are combined by stacked generalisation [29]:

$$z(t) = w_lex\ s_lex(t) + w_ex\ s_ex(t) + w_kb\ s_kb(t) + b, \quad P(fake \mid t) = \sigma(z(t)). \quad (3)$$

To avoid the optimistic bias of fitting the stacker on channel outputs produced for the same reviews the channels were trained on, the stacker weights are estimated on out-of-fold channel outputs: the training split is partitioned into K inner folds, each channel is fitted on $K - 1$ folds and applied to the held-out fold, and the stacker is fitted on the resulting out-of-fold scores. The channels are then refitted on the full training split for inference. The stacker has only four parameters, which keeps it stable for small training sets, and its weights show how much each evidence source is trusted in a given domain. Because (3) is additive, the contribution of each channel to a verdict is exactly

w_c s_c(t), and the explanation reports the full decomposition z = w_lex s_lex + w_ex s_ex + w_kb s_kb + b.

F. Grounded Explanation Generation

The explanation for a review is a structured object rendered into a four-part rationale whose steps follow the category structure of the knowledge base (Section III-C) in the step-by-step form of chain-of-thought reasoning [11]. (1) *Linguistic evidence*: the top-5 n-grams present in the review that push toward the verdict, each quoted verbatim. (2) *Knowledge indicators*: up to four triggered indicators with their descriptions, direction and z-scores. (3) *Precedents*: the number of fake reviews among the k retrieved exemplars and the closest exemplar with its label and similarity. (4) *Decision*: the verdict, the confidence max(P, 1 − P) and the additive decomposition of the log-odds. Every item is a span of the review, a measured property of it, or a retrieved labelled review, so the rationale contains no claim that cannot be checked against the input or the index. Fig. 3 shows rationales produced for a deceptive and a genuine review.

Table 1THE EXECUTABLE FRAUD-INDICATOR KNOWLEDGE BASE (25 ITEMS, THREE CATEGORIES; † IMPLEMENTS A KNOWLEDGE ITEM OF THE DESIGN ADOPTED IN SECTION III-C)

ID	Category	Indicator (measured on the review)	ID	Category	Indicator (measured on the review)
L1†	linguistic	density of superlatives / intensifiers	L14	linguistic	all-caps words
L2	linguistic	use of exclamation marks	L15	linguistic	rhetorical questions
L3	linguistic	first-person singular pronoun rate	L16	linguistic	repeated punctuation (!!, ...)
L4	linguistic	second-person (reader-directed) pronoun rate	L17†	linguistic	concrete spatial / sensory details
L5†	linguistic	generic promotional phrases	L18	linguistic	past-tense narration of experience
L6†	linguistic	numeric / quantitative details	L19	linguistic	function-word rate
L7†	linguistic	named, specific entities	S1	sentiment	sentiment extremity
L8	linguistic	lexical diversity	S2	sentiment	share of positive wording
L9	linguistic	average word length	S3	sentiment	share of negative wording
L10	linguistic	review length	S4	sentiment	balanced (mixed) sentiment
L11	linguistic	sentence-length variability	B1†	behavioural	extreme star rating
L12	linguistic	repeated phrasing	B2†	behavioural	mismatch between star rating and text sentiment
L13	linguistic	adverb (-ly) usage			

The union of the cited n-grams and the surface tokens that realise the triggered lexicon-based indicators defines the *evidence span set* E(t). This set is what the faithfulness evaluation in Section **VI removes from, or retains in, the review.**

G. Complexity and Updatability

Training fits two logistic models and builds the retrieval index, which takes seconds to minutes on a CPU; inference needs one encoder pass, one sparse and one dense scoring pass over the index (O(N) for N indexed reviews, trivially batched) and two sparse dot products. Because the index is non-parametric, newly labelled reviews can be appended to it at any time: the precedents shown in explanations immediately reflect them, while the stacker and the other channels remain unchanged. Section V-C shows that appending reviews immediately improves the retrieved precedents, while detection gains come from refitting the two logistic channels, which takes seconds to minutes.

IV. Experimental Setup

A. Datasets

We evaluate on two public benchmarks that represent the two dominant forms of review fraud. *OpSpam* is the Deceptive Opinion Spam corpus of Ott et al. [4], [30]: 1,600 reviews of 20 Chicago hotels, half truthful (collected from TripAdvisor and other platforms) and half deceptive (written on request by crowd workers), balanced across positive and negative polarity. Human judges perform near chance on it. *Amazon-CG* is the fake-reviews dataset of Salminen et al. [18]: 40,432 product reviews from ten Amazon categories, half original (OR) and half generated by a GPT-2 model fine-tuned on the originals (CG), each with its star rating. The corpora differ in size by a factor of 25, in review length (149 vs. 67 words on average) and in the nature of the fake class, which lets us test whether the same evidence channels apply to human and machine deception. Both are balanced, so accuracy is meaningful.

B. Baselines

We compare GRADE with representatives of each family of detector that can be trained and run under the same CPU budget. *Lexical models*: TF-IDF word uni/bigrams with multinomial naive Bayes, a linear SVM (C = 1) and logistic regression (C = 10), the strongest in-domain detectors reported for OpSpam [4]. *Feature-based models*: a random forest (300 trees) and a logistic regression on the 25 knowledge-base indicators, i.e., the hand-crafted linguistic/behavioural feature approach of early work [5]. *Pretrained-encoder models*: logistic regression on the frozen MiniLM sentence embedding. *Retrieval-only models*: the similarity-weighted k-NN vote of (1) with dense, BM25 and hybrid retrieval (k = 10), i.e., GRADE's exemplar channel used on its own. All models are trained on identical folds with identical features and seeds.

C. Metrics and Protocol

We report accuracy, precision, recall and F1 for the fake class together with the threshold-free AUC-ROC and AUC-PR. All detection results are 5-fold stratified cross-validation means with standard deviations (seed 42); the stacker of GRADE is fitted on out-of-fold channel outputs within each training fold (K = 5 inner folds on OpSpam, K = 2 on Amazon-CG), so no test review influences any component. Two further elements of the experimental design are drawn from KE-MLLM [19]. First, we reproduce its similarity-based knowledge retrieval as an alternative selection rule for the knowledge channel: the review and natural-language statements of the 25 indicators are embedded with MiniLM, and only the five items with the highest cosine similarity, the retrieval depth of the original configuration, are passed to the channel; a label-free control instead keeps the five indicators with the largest measured deviation $|z_j|$. Second, we employ its threshold-based faithfulness score, the fraction of reviews whose confidence in the predicted class drops by more than $\tau = 0.2$ when the cited evidence is removed, and report it together with ERASER's comprehensiveness and sufficiency [23] and the verdict flip rate (Section VI).

D. Implementation Details

The dense encoder is the ONNX export of all-MiniLM-L6-v2 (384 dimensions, maximum 256 word pieces for OpSpam and 128 for Amazon-CG). The exemplar channel retrieves k = 10 neighbours unless stated otherwise. The lexical channel uses sub-linear TF-IDF over word uni/bigrams with minimum document frequency 2 and L2-regularised logistic regression (C = 10, liblinear). The knowledge channel standardises the 25 indicators with training statistics and fits an L2-regularised logistic model (C = 1), as does the stacker. Sentiment indicators use the VADER lexicon [28]. All experiments run on one CPU core with 4 GB of memory; no GPU is used at any stage.

V. RESULTS

A. Detection Performance

Table II reports the main comparison. On OpSpam, GRADE reaches 90.0% accuracy and 90.0% F1 with an AUC-ROC of 96.7% and an AUC-PR of 96.9%. The strongest baseline is the lexical SVM (89.9% accuracy, 90.0% F1, 96.4% AUC-ROC): GRADE matches it on the thresholded metrics, improves both area metrics by 0.3 points. On Amazon-CG, GRADE improves on every baseline on every metric, reaching 95.2% accuracy and F1 (+0.4 and +0.5 over the SVM, +0.7 over logistic regression) and 99.0% AUC-ROC. The gains on Amazon-CG are consistent: GRADE exceeds the best baseline in all five folds (by 0.3 to 0.8 F1 points). On OpSpam, GRADE is on par with the strongest lexical models in-domain (fold-level differences within one point) while adding grounded explanations and a clear advantage when labelled data are scarce (Section V-C). The retrieval-only and knowledge-only detectors are clearly weaker than the lexical models in isolation (82.7% and 75.3% F1 on OpSpam; 82.1% and 82.7% on Amazon-CG): precedent retrieval and fraud indicators are complementary sources of evidence, not replacements for a text model.

Table 2Table 2DETECTION PERFORMANCE (5-FOLD CROSS-VALIDATION, MEAN ± STD, %). POSITIVE CLASS = FAKE REVIEW.

	OpSpam (hotel, human-written)						Amazon-CG (machine-generated)					
Method	**Acc**	**P**	**R**	**F1**	**AUC-ROC**	**AUC-PR**	**Acc**	**P**	**R**	**F1**	**AUC-ROC**	**AUC-PR**
TF-IDF + NB	87.6 ±0.5	82.9 ±0.5	94.8 ±1.3	88.4 ±0.5	96.2 ±0.4	96.4 ±0.4	89.2 ±0.5	84.5 ±0.7	96.0 ±0.2	89.9 ±0.4	97.6 ±0.1	97.9 ±0.1
TF-IDF + SVM	89.9 ±1.0	89.3 ±1.0	90.8 ±1.7	90.0 ±1.0	96.4 ±0.3	96.6 ±0.3	94.8 ±0.1	95.5 ±0.3	94.0 ±0.2	94.7 ±0.1	98.8 ±0.0	98.9 ±0.0
TF-IDF + LR	89.8 ±1.1	89.3 ±1.4	90.4 ±1.6	89.8 ±1.1	96.4 ±0.3	96.6 ±0.3	94.5 ±0.1	95.4 ±0.2	93.6 ±0.2	94.5 ±0.1	98.7 ±0.0	98.8 ±0.0
KB features + RF	74.7 ±2.3	74.7 ±2.3	74.6 ±2.7	74.7 ±2.4	81.7 ±2.8	79.6 ±3.5	85.7 ±0.4	84.6 ±0.6	87.2 ±0.2	85.9 ±0.3	93.7 ±0.2	93.6 ±0.3
KB indicators + LR	74.9 ±3.0	74.3 ±3.4	76.5 ±2.5	75.3 ±2.7	82.3 ±2.2	80.9 ±3.4	82.6 ±0.4	82.4 ±0.6	82.9 ±0.4	82.7 ±0.4	91.3 ±0.3	91.0 ±0.3
MiniLM + LR	81.9 ±1.2	81.1 ±0.8	83.2 ±2.1	82.2 ±1.3	90.4 ±1.5	90.5 ±1.1	80.3 ±0.3	80.4 ±0.4	80.1 ±1.0	80.2 ±0.4	88.3 ±0.3	88.1 ±0.4
kNN vote (dense, k=10)	71.7 ±1.5	66.3 ±1.4	88.5 ±3.0	75.8 ±1.4	79.9 ±1.8	77.2 ±2.5	67.3 ±0.4	62.3 ±0.3	87.8 ±0.5	72.9 ±0.3	78.9 ±0.3	78.4 ±0.7
kNN vote (BM25, k=10)	81.9 ±1.5	79.1 ±2.1	86.8 ±1.2	82.7 ±1.3	89.8 ±1.1	89.6 ±1.4	79.0 ±0.4	71.7 ±0.4	96.0 ±0.3	82.1 ±0.3	92.9 ±0.1	91.8 ±0.3
kNN vote (hybrid, k=10)	79.5 ±1.8	73.5 ±1.6	92.4 ±1.7	81.8 ±1.6	89.6 ±0.8	88.5 ±1.5	73.8 ±0.3	66.9 ±0.3	94.2 ±0.3	78.2 ±0.2	88.3 ±0.2	87.0 ±0.6
GRADE (ours)	**90.0 ±0.8**	**90.1 ±1.0**	**89.9 ±1.2**	**90.0 ±0.8**	**96.7 ±0.2**	**96.9 ±0.2**	**95.2 ±0.2**	**95.6 ±0.3**	**94.8 ±0.2**	**95.2 ±0.2**	**99.0 ±0.0**	**99.1 ±0.0**

B. Ablation of Evidence Channels, Retrieval and Knowledge Selection

Table III removes one channel at a time. The lexical channel is the backbone in both domains: without it, F1 drops to 85.4% on OpSpam and 87.9% on Amazon-CG. The knowledge channel contributes most on Amazon-CG, where removing it costs 0.7 F1 and 0.25 AUC-ROC points and where the stacker weights it (0.48) above the exemplar channel (0.30), consistent with the strong stylistic regularities of generated text. On OpSpam the knowledge channel is weighted lightly (0.15), while the exemplar channel receives a weight of 1.41 and raises AUC-ROC from 96.3% to 96.7%. In-domain on Amazon-CG the exemplar channel leaves the thresholded metrics unchanged, since the lexical model already separates the classes almost perfectly; its benefits appear in the low-data setting below and, above all, in the explanations, which gain concrete precedents. The retrieval rows of Table III compare dense, sparse and hybrid retrieval. Dense retrieval alone and BM25 alone yield very similar detection quality once stacked (90.1%/90.5% F1 on OpSpam, 95.2%/95.3% on Amazon-CG), and the hybrid achieves the best AUC-ROC on

OpSpam (96.7%). As standalone voters, BM25 is far stronger than dense cosine (82.7% vs. 75.8% F1 on OpSpam; 82.1% vs. 72.9% on Amazon-CG): deceptive reviews reuse distinctive phrases, which exact matching captures, whereas semantic similarity retrieves reviews about the same hotel or product regardless of label. The hybrid inherits the precision of BM25 while adding paraphrase-level neighbours, and we use it by default because it exposes both kinds of precedent.

Table 3ABLATION OF EVIDENCE CHANNELS, RETRIEVAL MODES AND KNOWLEDGE-SELECTION RULES (5-FOLD MEAN, %; AUC = AUC-ROC)

	OpSpam			Amazon-CG		
Variant	**Acc**	**F1**	**AUC**	**Acc**	**F1**	**AUC**
GRADE (full)	**90.0**	**90.0**	**96.7**	**95.2**	**95.2**	**99.0**
w/o lexical channel	85.3	85.4	91.9	87.8	87.9	95.1
w/o exemplar channel	89.9	89.9	96.3	95.3	95.2	99.0
w/o knowledge channel	89.9	89.9	96.7	94.6	94.5	98.8
lexical only	89.6	89.5	96.4	94.6	94.5	98.7
exemplars only	81.8	82.6	89.6	80.4	80.8	88.3
knowledge only	74.8	75.3	82.3	82.6	82.7	91.3
retrieval: dense only	90.1	90.1	96.4	95.2	95.2	99.0
retrieval: BM25 only	90.5	90.5	96.5	95.3	95.3	99.0
retrieval: hybrid (RRF)	90.0	90.0	96.7	95.2	95.2	99.0
knowledge only, similarity top-5	64.3	66.1	68.9	64.0	63.5	69.8
knowledge only, deviation top-5	70.1	71.1	76.8	78.5	78.6	87.0
GRADE, similarity top-5 knowledge	90.1	90.1	96.7	94.6	94.6	98.8
GRADE, deviation top-5 knowledge	90.0	89.9	96.6	95.1	95.1	98.9

The last four rows of Table III compare knowledge-selection rules. When the knowledge channel sees only the five items whose descriptions are most similar to the review, it reaches 66.1% F1 and 68.9% AUC-ROC on OpSpam and 63.5% and 69.8% on Amazon-CG, compared with 75.3%/82.3% and 82.7%/91.3% when all calibrated indicators are measured; the deviation-based control lies in between (71.1% and 78.6% F1). The retrieved items explain the gap: on hotel reviews the similarity rule selects the spatial-detail item for 99% and the past-tense item for 89% of the reviews, because their descriptions match the topic of almost every review, and on average only 1.4 (OpSpam) and 1.2 (Amazon-CG) of its five items coincide with the five indicators that deviate most. Measured against the reviews themselves, 36% and 32% of the similarity-selected items are exhibited in the direction of the verdict, against 56% and 61% for the deviation-based control. Similarity retrieval suits prompting, where the LLM itself judges whether a retrieved pattern is present; executable knowledge makes that judgement explicit and measurable, which turns the surfaced indicators into reliable evidence. In the full model the stacker down-weights a weaker knowledge channel, so the three rules perform alike on OpSpam (90.0%, 90.1% and 89.9% F1), whereas on Amazon-CG, where knowledge carries more weight, similarity-based selection lowers F1 from 95.2% to 94.6% and AUC-ROC from 99.0% to 98.8%. Varying k from 1 to 50 (5-fold means) shows that GRADE is stable over the whole range: F1 stays within 90.0–90.3% on OpSpam and 95.2–95.3% on Amazon-CG, and AUC-ROC varies by at most 0.3 points; the stacker adapts the exemplar weight to k (from 0.26 at k = 1 to 1.97 at k = 50 on OpSpam) as the vote becomes smoother. We use k = 10 as a balance between evidence breadth and the length of the precedent list shown in explanations.

C. Data Efficiency and Index Growth

Retrieval augmentation uses every labelled review directly as evidence, which favours small labelled sets. Fig. 2 trains all components on 5% to 100% of a training fold. The advantage of GRADE over the lexical model is largest when data are scarce: with 128 OpSpam reviews GRADE reaches 79.2% F1 versus 73.9% for TF-IDF + LR (+5.3 points), with 320 reviews 86.5% versus 83.9%, and with 640 reviews 89.4% versus 87.7%; on Amazon-CG the margin is 1.2–1.3 points from 1,617 to 3,234 reviews and shrinks to 0.9 at 16k reviews. The k-NN vote alone is weak in all regimes, so the gain comes from stacking precedents and indicators on top of the lexical score.

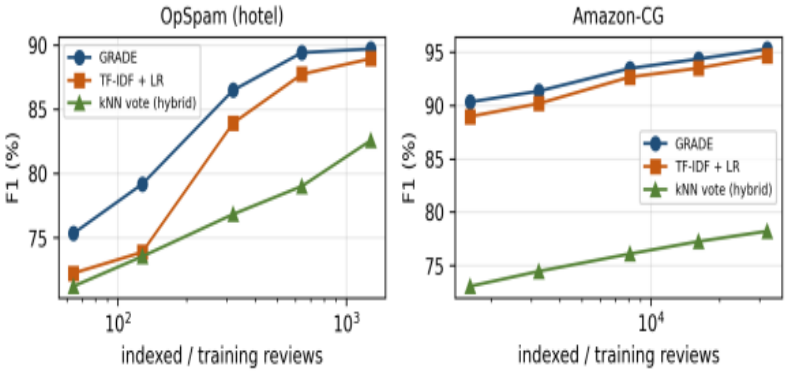


Figure 2Data efficiency: F1 on the held-out fold as a function of the number of indexed/training reviews (log scale).

Newly labelled reviews can be appended to the index immediately or used to refit the channels. In a fold-1 experiment in which GRADE is first trained on half of the training fold, appending the other half to the index improves the evidence shown to the user at once: the similarity of the closest precedent rises from 0.699 to 0.715 on OpSpam and from 0.591 to 0.615 on Amazon-CG, and the share of precedents carrying the true label from 65.8% to 67.4% and from 68.1% to 69.8%. Detection gains come from refitting the lexical and knowledge channels (F1 from 87.4% to 88.4% and from 94.4% to 95.3%), which takes only 1.5 s and 99 s, respectively (Section V-D).

D. Computational Cost

On one CPU core, training GRADE on an OpSpam fold, including the inner cross-validation needed for the stacker, takes 1.5 s; on an Amazon-CG fold with 32,345 indexed reviews it takes 99 s. End-to-end inference, including both retrievals and the rendering of the explanation, takes 54 ms per review on OpSpam and 150 ms on Amazon-CG, where scoring against the larger index dominates and batching helps further. GRADE thus brings explainable detection into the range of commodity hardware, with 23k (OpSpam) and 140k (Amazon-CG) trained parameters, essentially all of them interpretable n-gram weights.

VI. Explanation Faithfulness

A. Protocol

An explanation is faithful if the evidence it cites is what the model actually used. We test this with the perturbation protocol of Section IV-C. For 100 randomly sampled test reviews per dataset (fold-1 models), we delete the evidence span set E(t) cited by GRADE from the review and re-run the *entire* pipeline on the perturbed text, including re-encoding and re-retrieval. Let $c(t) = P(\hat{y} \mid t)$ be the confidence in the originally predicted class $\hat{y}$. Comprehensiveness is $c(t) - c(t \setminus E)$, sufficiency is $c(t) - c(E)$ with only the cited spans kept, faithfulness@0.2 is the fraction of reviews with comprehensiveness above $\tau = 0.2$, and the flip rate is the fraction of changed verdicts. All methods are given the *same token budget*: the number of distinct tokens that GRADE cites for a review is used as the budget for the competing explanations of the same review. We compare four explanations of the same GRADE model: the full GRADE evidence set; GRADE's lexical terms only; LIME [22] run on the full pipeline as a black box (400 samples per review); and a random

selection of the same number of word tokens, which calibrates the effect of any deletion of that size.

Table 4EXPLANATION FAITHFULNESS ON 100 HELD-OUT REVIEWS PER DATASET AT THE SAME TOKEN BUDGET (COMPR. = COMPREHENSIVENESS, SUFF. = SUFFICIENCY; FAITH@0.2, FLIP AND TOKENS REMOVED IN %)

Explanation (same budget)	Compr.	Suff.	Faith@0.2	Flip	Tokens
OpSpam					
GRADE evidence (full)	0.303	-0.097	58.0	35.0	9.8%
GRADE lexical terms only	0.271	-0.097	52.0	33.0	6.8%
LIME (400 samples)	0.311	-0.094	60.0	38.0	7.9%
Random tokens	0.033	0.255	7.0	12.0	9.5%
Amazon-CG					
GRADE evidence (full)	0.282	-0.053	44.0	30.0	23.0%
GRADE lexical terms only	0.264	-0.054	41.0	28.0	19.6%
LIME (400 samples)	0.376	-0.027	56.0	43.0	17.6%
Random tokens	0.095	0.303	19.0	15.0	21.5%

B. Results

Table IV reports the results. Removing the evidence cited by GRADE, which covers 9.8% of the tokens of an OpSpam review and 23% of the tokens of the much shorter Amazon-CG reviews, lowers the confidence in the predicted class by 0.30 and 0.28 on average, flips 35% and 30% of the verdicts, and exceeds the $\tau = 0.2$ threshold for 58% and 44% of the reviews. The random control at the same budget lowers confidence by only 0.03 and 0.10, flips 12% and 15% of the verdicts and passes the threshold for 7% and 19%: GRADE's evidence is three to nine times more comprehensive than an arbitrary deletion of the same size, so the drops do not merely reflect shortened inputs. Sufficiency is negative for GRADE on both datasets (−0.097 and −0.053): keeping only the cited spans produces a prediction that is at least as confident as the prediction on the full review, i.e., the cited evidence alone carries the decision, whereas keeping random tokens loses 0.25 and 0.30 of confidence. LIME, which is given 400 additional evaluations of the full pipeline per review to fit a local surrogate, reaches a comprehensiveness of 0.31 on OpSpam, on par with GRADE's 0.30, and 0.38 on Amazon-CG, where it finds somewhat more disruptive tokens; GRADE's evidence, in turn, is at least as sufficient on both datasets (−0.097 vs. −0.094 and −0.053 vs. −0.027; lower is better). GRADE's rationale is thus competitive with a dedicated post hoc explainer while being a free by-product of the decision, and it additionally reports precedents and indicator deviations that no token-attribution method can express. Restricting the evidence to the lexical terms (second row) lowers comprehensiveness by 0.02–0.03, so the indicator tokens carry signal of their own.

VII. Transfer Across Polarity, Category and Dataset

Table V evaluates models trained in one condition and applied without any adaptation to another. We consider a *polarity shift* within OpSpam (train on the 800 positive reviews, test on the 800 negative ones, and vice versa), a *category shift* within Amazon-CG (train on eight product

Table 5Table 5Table 5TRANSFER ACROSS POLARITY, PRODUCT CATEGORY AND DATASET (F1 / AUC-ROC, %)

Transfer setting	n_train / n_test	TF-IDF+LR	MiniLM+LR	KB+LR	kNN (hybrid)	GRADE w/o KB	GRADE
OpSpam: positive → negative	800 / 800	67.3 / 85.6	47.9 / 80.3	32.3 / 74.1	67.7 / 81.0	67.1 / 87.0	64.2 / 87.5
OpSpam: negative → positive	800 / 800	76.7 / 91.4	63.3 / 79.7	61.8 / 75.6	72.5 / 79.0	74.0 / 91.4	73.4 / 91.4
Amazon-CG: held-out Books, Kindle	31332 / 9100	76.8 / 95.9	73.5 / 85.1	84.7 / 93.1	74.9 / 82.1	78.8 / 96.0	85.5 / 97.6
Amazon-CG: held-out Electronics, Tools	32586 / 7846	92.5 / 98.4	71.5 / 82.7	81.5 / 90.7	77.6 / 84.8	92.6 / 98.4	93.6 / 98.7
OpSpam → Amazon-CG (cross-dataset)	1600 / 40432	13.1 / 41.1	27.3 / 55.7	66.2 / 68.4	45.5 / 50.2	13.0 / 42.6	17.9 / 46.5
Amazon-CG → OpSpam (cross-dataset)	40432 / 1600	0.0 / 50.6	12.6 / 59.0	12.1 / 63.1	32.8 / 50.8	0.0 / 50.7	0.0 / 55.3

categories, test on two held-out ones) and the *cross-dataset shift* between human-written hotel deception and machine-generated product reviews.

Under category shift, GRADE transfers markedly better than the text-only models. With Books and Kindle held out it reaches 85.5% F1 and 97.6% AUC-ROC against 76.8% and 95.9% for TF-IDF + LR, a gain of 8.7 F1 points, and with Electronics and Tools held out it reaches 93.6% against 92.5%. The ablation column shows where the robustness comes from: without the knowledge channel GRADE drops to 78.8% on the Books/Kindle split, and the knowledge indicators alone reach 84.7%. Generated reviews share stylistic regularities (sentiment extremity, short length, little concrete detail, rating-text alignment) that do not depend on the product vocabulary; the calibrated indicators capture exactly these, whereas n-gram models partly learn category-specific wording, and the stacker inherits the robustness through the weight it assigns to the knowledge channel. Self-play contrastive fine-tuning for web search agents further illustrates the value of testing retrieval-oriented systems under changing input conditions and distribution shifts [32].

Under polarity shift, GRADE preserves its ranking quality while the decision threshold moves. GRADE's ranking quality is preserved or improved (AUC-ROC 87.5% vs. 85.6% for the lexical model in the positive→negative direction and 91.4% in both cases in the reverse direction), but its thresholded F1 is lower (64.2% vs. 67.3% and 73.4% vs. 76.7%), and the ablation attributes most of the gap in the positive→negative direction to the knowledge channel: several sentiment indicators that are calibrated on positive reviews reverse their meaning on negative ones, which shifts the decision threshold even though the ordering of reviews remains good. Re-calibrating the indicator statistics on a small unlabelled target sample addresses this, and the additive model identifies the responsible channel directly from its contribution.

The cross-dataset setting is a stress test between two phenomena that share little surface signal, crowd-sourced hotel deception and GPT-2 product reviews. Here all models stay between 41% and 68% AUC-ROC and the text models default to the majority style of their source domain, while the knowledge channel retains the most signal (68.4% and 63.1% AUC-ROC): a small core of indicators, notably sentiment extremity, generic phrasing and the absence of concrete details, is shared across both forms of fraud, whereas lexical and precedent evidence is domain-bound.

VIII. Qualitative and Error Analysis

Fig. 3 shows two rationales produced for OpSpam test reviews, one deceptive and one genuine. The deceptive review (Example 1) is a short, generic complaint about a named hotel. The lexical channel cites "intercontinental chicago", "chicago hotel" and "the food": crowd-sourced deceptive reviews in OpSpam repeatedly name the hotel and the city, a pattern that genuine guests, who take the setting for granted, rarely exhibit, and the model has learned it as a strong cue. All ten retrieved precedents are deceptive, and the exemplar channel contributes +1.41 to a log-odds of +3.94. The genuine review (Example 2) is classified with similar confidence for the opposite reasons: it contains concrete, checkable detail ("3 blocks", "2 blocks", "ten minutes", the concierge's name), which the numeric-detail indicator flags at z = +1.6, it uses few first-person singular pronouns and no exclamation marks, and only three of its ten precedents are deceptive. Every claim in these rationales can be checked against the review or the index, and disagreement between channels is exposed rather than hidden.

Example 1 — ground truth: fake

Review: "The InterContinental Chicago hotel terrible. First of all, the continental breakfast was a joke. The food was unhealthy and unsatisfactory. After our first night, we expected the maid to come clean up our room. We came back from site-seeing to find that our ro…"

GRADE rationale: Verdict: DECEPTIVE (confidence 0.98). (1) Linguistic evidence: the phrase(s) "chicago", "chicago hotel", "the food", "intercontinental chicago", "food" are characteristic of deceptive reviews in the indexed corpus. (2) Knowledge indicators: low numeric / quantitative details (z = -0.7); low sentiment extremity (z = -2.0); high share of negative wording (z = +1.0); high function-word rate (z = +0.8). (3) Precedents: 10 of the 10 most similar indexed reviews are deceptive (closest: "Me and my wife stayed at the Omni hotel in Chicago for a customer training at a …", deceptive, sim = 0.67). (4) Decision: log-odds = lexical +2.92 + exemplar +1.41 + knowledge -0.10 + bias -0.29 = +3.94.

Example 2 — ground truth: genuine

Review: "Stayed at this hotel with 3 friends or 4 nights. The hotel was clean and tidy, throughout. The Concierge Christopher was excellent and helped as with all our needs, gave us discount vouchers etc. Hotel was in excellent position. 3 blocks from John Hancok Build…"

GRADE rationale: Verdict: GENUINE (confidence 0.97). (1) Linguistic evidence: the phrase(s) "blocks", "excellent", "concierge", "we", "etc" are characteristic of genuine reviews in the indexed corpus. (2) Knowledge indicators: high numeric / quantitative details (z = +1.6); low first-person singular pronoun rate (z = -0.9); low function-word rate (z = -0.9); low use of exclamation marks (z = -0.6). (3) Precedents: 3 of the 10 most similar indexed reviews are deceptive (closest: "We got an outstanding rate by signing up for the hotels' website newsletter. Aff…", genuine, sim = 0.70). (4) Decision: log-odds = lexical -1.87 + exemplar -0.56 + knowledge -0.85 + bias -0.29 = -3.57.

Fig. 3. Rationales generated by GRADE for a deceptive and a genuine OpSpam review (fold-1 model). Every cited item is a span of the review, a measured indicator or a retrieved labelled precedent, and the final line shows the additive decomposition of the log-odds.

Finally, we analyse the errors of the fold-1 models. GRADE misclassifies 37 of 320 OpSpam reviews and 378 of 8,087 Amazon-CG reviews, with false positives and false negatives in roughly equal numbers. The additive design makes errors easy to characterise: in 68% of the OpSpam errors and 76% of the Amazon-CG errors the three channels disagree on the sign of the verdict, against 36% of the correctly classified reviews in both datasets, and the error rate is 6.2% and 1.8% when all three channels agree versus 19.8% and 9.4% when they do not. Errors are also made with visibly lower confidence (mean 0.71 and 0.74 versus 0.93 and 0.96 for correct verdicts) and, on Amazon-CG, concentrate on very short reviews (median 18 words versus 41). Channel

disagreement and low confidence are therefore reliable, and transparent, triggers for routing a review to a human moderator.

IX. Conclusion

This paper presented GRADE, a retrieval-augmented framework for explainable fake review detection in which every verdict is grounded in retrieved labelled precedents, calibrated fraud indicators and lexical cues, combined by an additive stacker whose log-odds decompose exactly into channel contributions. On two public benchmarks covering human-written and machine-generated fraud, GRADE matches or exceeds the strongest lexical detectors in-domain (90.0% and 95.2% accuracy; 96.7% and 99.0% AUC-ROC), is substantially more accurate when labelled data are scarce, transfers across product categories far better than text-only models, and can be refitted on newly labelled reviews in seconds to minutes. Its explanations are faithful by construction: in the perturbation tests used for LLM-based detectors they are competitive with a dedicated LIME surrogate at no additional cost and exceed random evidence several-fold, and the whole system runs in tens of milliseconds per review on a single CPU core. These results show that the explainability goals of knowledge-enhanced LLM detection can also be met by grounded retrieval without parameter updates to a large model, and that measuring fraud knowledge, rather than retrieving it by textual similarity, yields a markedly stronger knowledge channel. The two directions are complementary: the evidence GRADE assembles (precedents with labels and similarities, measured indicators with deviation scores, decisive phrases) is precisely the context a chain-of-thought LLM needs to produce rationales that are both fluent and verifiable, which makes GRADE a natural evidence layer for LLM-based moderation pipelines. Because the public corpora expose only star ratings as metadata, the behavioural category was instantiated with rating-based indicators, and polarity-specific calibration of the indicator statistics can further stabilise the decision threshold across sentiment polarities. Measuring burst and account-age red flags on platform data, re-calibrating indicator statistics on unlabelled target reviews and evaluating the precedent-based explanations with moderators are the corresponding next steps.

References

M. Luca and G. Zervas, “Fake it till you make it: Reputation, competition, and Yelp review fraud,” Management Science, vol. 62, no. 12, pp. 3412–3427, 2016.

Y. Wu, E. W. T. Ngai, P. Wu, and C. Wu, “Fake online reviews: Literature review, synthesis, and directions for future research,” Decision Support Systems, vol. 132, p. 113280, 2020.

R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, “Fake reviews detection: A survey,” IEEE Access, vol. 9, pp. 65771–65802, 2021.

M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2011, pp. 309–319.

J. Wang, Y. Ge, J. Liu, and B. Jiang, “Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms,” Journal of Trends in Finance and Economics, vol. 1, no. 1, pp. 85–100, 2024.

S. Rayana and L. Akoglu, “Collective opinion spam detection: Bridging review networks and metadata,” in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2015, pp. 985–994.

Y. Ren and D. Ji, “Neural networks for deceptive opinion spam detection: An empirical study,” Information Sciences, vol. 385–386, pp. 213–224, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, “Enhancing graph neural network-based fraud detectors against camouflaged fraudsters,” in Proc. 29th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), 2020, pp. 315–324.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in Proc. Int. Conf. Learn. Representations (ICLR), 2022.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 24824–24837.

S. Jain and B. C. Wallace, "Attention is not explanation," in Proc. NAACL-HLT, 2019, pp. 3543–3556.

A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 4198–4205.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.

O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, "In-context retrieval-augmented language models," Transactions of the Association for Computational Linguistics, vol. 11, pp. 1316–1331, 2023.

D. Teng, "TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows," World Journal of Engineering Research, vol. 3, no. 6, pp. 55–66, 2025.

J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen, "What makes good in-context examples for GPT-3?" in Proc. Deep Learning Inside Out (DeeLIO 2022): 3rd Workshop Knowl. Extraction Integration Deep Learn. Archit., 2022, pp. 100–114.

J. Salminen, C. Kandpal, A. M. Kamel, S.-g. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," Journal of Retailing and Consumer Services, vol. 64, p. 102771, 2022.

J. Chen, J. Liu, Y. Liang, and M. Zhou, "KE-MLLM: A knowledge-enhanced multi-sensor learning framework for explainable fake review detection," Applied Sciences, vol. 16, no. 6, p. 2909, 2026.

K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: Retrieval-augmented language model pre-training," in Proc. 37th Int. Conf. Mach. Learn. (ICML), PMLR vol. 119, 2020, pp. 3929–3938.

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD), 2016, pp. 1135–1144.

J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 4443–4458.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 5776–5788.

N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992.

S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2009, pp. 758–759.

C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proc. 8th Int. AAAI Conf. Weblogs Social Media (ICWSM), 2014, pp. 216–225.

D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technologies (NAACL-HLT), 2013, pp. 497–501.

F. Zhou, “Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders,” Innovation and Technology Studies, vol. 3, no. 1, pp. 27–41, 2026.

M. Rhee et al., “Enhancing Web Search Agents With Self-Play Contrastive Fine-Tuning,” IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3717594.